DOCUMENT RESUME

ED 372 117 TM 021 958

AUTHOR Fan, Xitao; Mathews, Tom A.
TITLE Using Bootstrap Procedures To Assess the Issue of Predictive Bias in College GPA Prediction for Ethnic Groups.
PUB DATE Apr 94
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Bias; Class Rank; College Students; Ethnic Groups; *Grade Point Average; Higher Education; High Schools; *Prediction; Predictor Variables; *Scores; *Statistical Significance; Statistical Studies
IDENTIFIERS *Bootstrap Methods; Empirical Research; *Scholastic Aptitude Test

ABSTRACT

In this paper the bootstrap technique was applied to the evaluation of potential predictive bias for different ethnic groups using the Scholastic Aptitude Test (SAT) and high school ranks to predict college grade point averages (GPAs). Data of three ethnic groups were used; the total number of subjects was close to 5,000. Both conventional statistical-significance testing and the bootstrap procedure were used, and the results from the two approaches compared. For the data analyzed, when one predictor (SAT score) was used for GPA prediction, the two approaches provided somewhat different results. When two predictors (SAT score and high school rank) were used, the two approaches provided similar results. Where difference occurred, it was suggested that the results from the bootstrap approach might be more plausible. The paper demonstrated the feasibility of applying the bootstrap technique to the research in the area of predictive bias. Because the bootstrap procedure is more empirically grounded, some problems associated with conventional statistical-significance testing may be avoided. Three tables and four figures are included. (Contains 32 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
XITAO FAN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

ED 372 117

Using Bootstrap Procedures to Assess the Issue of Predictive Bias in College GPA Prediction for Ethnic Groups

Xitao Fan

Utah State University

Tom A. Mathews

Texas A&M University

Paper presented at the Annual Meeting of the American Educational Research Association, April 7, 1994. Session # 42.49.

SP 034 958

BEST COPY AVAILABLE

Abstract

In this paper, the bootstrap technique was applied to the evaluation of potential predictive bias for different ethnic groups for using SAT and high school rank to predict college GPA. Data of three ethnic groups were used, with the total number of subjects in the data being close to 5,000. Both the conventional statistical significance testing approach and the bootstrap procedure were used, and the results from the two approaches compared. For the data analyzed, when one predictor (SAT score) was used for GPA prediction, the two approaches provided somewhat different results. When two predictors (SAT score and high school rank) were used, the two approaches provided similar results. Where difference occurred, it was suggested that the results from the bootstrap approach might be more plausible. The paper demonstrated the feasibility of applying bootstrap technique to the research in the area of predictive bias investigation. Since bootstrap procedure is more empirically grounded, some problems associated with conventional statistical significance testing approach may be avoided.

Background

In psychological and educational measurement, the issue of differential predictive validity for different ethnic groups has been prominent for many years (Cleary, 1968; Hunter, Schmidt & Rauschenberger, 1984; Jensen, 1980; Reynolds, 1982). The prediction of college GPA, which is often the empirical foundation for making admission decisions, is no exception (Goldman & Richards, 1972; Hand & Prather, 1985; Sue & Abe, 1988;). Due to the socially sensitive nature of the predictive bias issue, it is imperative for higher education institutions to conduct empirical investigation and prove that their prediction of college GPA is not biased for different ethnic groups.

Statistically, the prediction of college GPA is implemented through regression analysis, with certain test scores and high school academic variables as the independent variables, and the college GPA as the dependent variable. In order to prove that predictive bias, especially bias against socially disadvantaged minority groups, does not exist, separate regression analyses for different ethnic groups can be conducted, and the resultant regression models compared. If an institution uses a common regression line for college GPA prediction, then the necessary condition for the absence of predictive bias is the existence of a common regression line for the ethnic groups. Since regression analysis is conducted separately for individual groups, this condition is satisfied only when the regression lines of the different groups share the same intercept and regression

coefficients. In other words, the absence of predictive bias requires that the different ethnic groups have the common intercept and slope(s).

When one predictor variable is used to predict performance on a criterion, three possible situations are depicted in Figure 1. In Figure 1(a), the two groups share the same regression line (same intercept and slope), and there is no predictive bias when this common regression line is used for prediction. In Figure 1(b), the two groups have the same regression slope, but different intercepts. In this situation, the use of a common regression line will systematically underpredict the performance of Group A members, and overpredict for Group B members. In Figure 1(c), both the intercepts and slopes of the regression lines of the two groups are different, a situation which can be properly treated by the Aptitude-Treatment-Interaction (ATI) model (Rogosa, 1980; Willson, 1985). In this situation, predictive bias exists for the two groups, but whether such bias results in over- or under-prediction depends on which score range the predictor score falls in.

Conventional statistical methods for investigating predictive bias deals with the situations depicted by Figure 1(a) and Figure 1(b). There are several ways to compare regression models for different groups in order to assess if predictive bias exists, and all of them involve statistical significance testing: 1) comparing intercept and slope separately (Kerlinger and Pedhazur, 1973; Pedhazur, 1982); 2) comparing intercepts and

slopes simultaneously (Potthoff, 1966; Reynolds, 1982); and 3) using analysis of variance (ANOVA) to analyze residuals of different groups after a common regression model is fitted to data of all groups as a whole (Reynolds, 1980, 1982).

Generally, the first two methods are applicable in situations in which only one predictor is involved. In this case, we are dealing with a regression line, and it is relatively straight-forward to compare intercepts and slopes. When more than one predictor is involved, we are no longer dealing with a regression line, but instead, with a regression plane. To assess the equality of two regression planes is no longer as straight-forward as for the case of a regression line (the planes can be more than three dimensional if we have more than three predictors). In this case, the ANOVA technique of comparing residuals of different groups can be effective for investigating the existence or absence of predictive bias. In this approach, a common regression plane is fitted to all the subjects without regard to ethnic group membership. If the prediction does not have equal prediction accuracy for different groups, or if systematic over- or under-prediction exists, the groups will differ in their residuals, and such difference can be detected through ANOVA analysis.

In educational and psychological research, the overreliance on statistical significance testing has been challenged on several grounds, including issues related to sample size and to the validity of the theoretical assumptions underlying parametric

statistical techniques (Carver, 1978; Thompson, 1989). The sample size issue becomes prominent due to the fact that, theoretically, any null hypothesis can be rejected (statistically significant) when sample size is large enough. This fact is often neglected in the interpretation of statistical significance, and as a result, the importance of statistical significance tends to be greatly exaggerated in research practice. As to the assumptions required for parametric statistical techniques, often, these assumptions are difficult, sometimes impossible, to assess or satisfy.

To avoid the blind reliance on theoretical sampling distributions, which are the basis of parametric statistical testing, researchers have turned to methods which are more empirically grounded. Bootstrap procedure, which is computing-intensive in nature, has become prominent in recent years as a complement to the traditional statistical significance testing, or an alternative approach to making statistical decisions (Thompson, 1993). Instead of relying on the theoretical sampling distribution of a statistic, bootstrap procedure, through repeated resampling with replacement from the original sample, generates estimated empirical sampling distribution for the statistic of our interest, upon which our statistical decisions can be based (Diaconis & Efron, 1983; Efron, 1979; Lunneborg, 1990; Thompson, 1992a). In this sense, bootstrap procedure attempts to avoid the pitfalls associated with the traditional statistical significance testing, such as problems related to

sample size issue, the overreliance on the correctness of the theoretical assumptions for our data in hand, etc. Since its debut in the late 70's (Efron, 1979), bootstrap method has gradually attracted the attention of the researchers in the educational and psychological research arena (Lunneborg, 1983; Lunneborg, 1987). Along with the easier access to powerful computing facilities, this method becomes more attractive than before.

Researchers in the educational and psychological research arena have applied the bootstrap technique to a variety of research problems, ranging from measurement issues, such as item discrimination and item bias indices, to multivariate statistical techniques, such as principal component analysis, factor analysis, and structural equation modeling (Bentler, 1992; Daniel, 1992; Harris & Kolen, 1988, 1989; Lambert, Wildt & Durand, 1990, 1991; Mendoza, Hart & Powell, 1991; Thompson, 1992a, 1992b). Some researchers also provided theoretical rationale or simulation results attesting to the applicability of this procedure to some widely used statistical methods (Bickel & Freedman, 1981; Freedman, 1981; Wu, 1986).

Bootstrap procedure may also be used as a viable alternative to statistical significance testing in the area of test predictive bias research. Instead of relying on sample sizes and theoretical sampling distributions for making statistical decisions, sampling distributions for the intercept and regression coefficients can be empirically estimated through

intensive bootstrap resampling, and the sample estimates from the separate regression analyses for different groups can be examined relative to the empirical distribution constructed through bootstrap resampling.

The purpose of this paper is to apply the bootstrap resampling procedure to investigation of predictive bias related to college GPA prediction for different ethnic groups. Three ethnic groups, the white, African-American and Hispanic-American, were involved in the study. Two predictors, test score of Scholastic Aptitude Test (SAT) and high school rank, were used both individually and jointly for the prediction of the first year cumulative Grade Point Average (GPA).

Methods

Data Source

The data of a freshmen cohort entering a major research university in the southwest of U.S. in 1990 were used for the study. The data contained about 6,000 students. Three broadly defined ethnic groups, White, African-American, and Hispanic-American, were used in the study. Two variables, SAT score and relative high school rank[1], were used as predictor variables, and the prediction for the first year cumulative GPA in the university was examined. The number of usable subjects for the

[1] Relative high school rank is calculated as:
$$Rank_{rel} = (1 - Rank/Class\ Size) \times 100$$
Thus the range of the $Rank_{rel}$ is 1 to 100, with 100 being the top and 1 the bottom.

three ethnic groups were 4248, 195, and 509 respectively for White, African-American and Hispanic-American groups.

Conventional Approaches

For the purpose of comparison, conventional significance testing approach was used to investigate potential predictive bias by comparing slopes and intercepts of the regression lines for the three ethnic groups. First, SAT score was used as sole predictor for the first year cumulative GPA. Separate regression lines were fitted for the three groups, and the slopes and intercepts of the resultant three regression lines were statistically tested using the procedures described in Kerlinger and Pedhazur (1973) and Pedhazur (1982).

In the second analysis, both SAT score and relative high school rank were used jointly for predicting the first year cumulative GPA. A common regression plane was fitted to all the subjects regardless of ethnic group membership. Potential predictive bias was investigated by applying ANOVA technique to test the equality o mean residuals for the ethnic groups. Unequal mean residuals for the different ethnic groups are indications of systematic under- or over-prediction for certain ethnic groups, thus indicating the existence of predictive bias.

Bootstrap Approach

Under the bootstrap approach, statistical decisions are based on the empirically estimated sampling distribution for the

statistic of our interest, rather than on the theoretical sampling distribution and the sample sizes. For the purpose of this study, the following were the two competing hypotheses:

H_0: For college GPA prediction, White, African-American and Hispanic-American groups could be treated as <u>one</u> population. Any differences in regression lines (planes, if more than one predictors) for the three groups would not exceed what could be expected from chance fluctuation.

H_1: The three groups were samples from <u>different</u> populations, and their regression lines (planes) would differ to such a degree that the use of a common regression line (plane) for GPA prediction would result in systematic under- or over-prediction for certain ethnic groups.

If we accepted the null hypothesis, we were basically saying that, for the purpose of GPA prediction, African-American and Hispanic-American groups were the same as the white group, thus they could be treated as samples from the same population. Since we were dealing with random samples from one population, any observed differences in GPA prediction for these three groups would not exceed what could be expected from random sampling variation. In order to determine which of the two competing hypotheses was tenable for our data, we would need to know:

1) a) When using SAT as the sole predictor, how much random sampling variation could be expected for the slope and

intercept of the prediction line for one population?

b) When using both SAT and the relative high school rank for GPA prediction, how much random sampling variation could be expected for the residuals for one population?

2) Did the observed differences among the three ethnic groups for regression lines (for the case of one predictor) or for the residuals (for the case of using two predictors) exceed what could be expected from random sampling variation?

If the answer to the last question is affirmative, we would conclude that the null hypothesis was not tenable, and this would be evidence to support the alternative hypothesis.

The empirical estimation for the sampling variation was achieved through bootstrap procedure. The white group was used as the reference group, and the sampling distributions for the statistics were empirically estimated by bootstrapping the white group data. Since the sampling distributions would be affected by sample sizes (In our data, African-American: n=195; Hispanic-American: n=509), for each statistic of our interest, sampling distributions were estimated both for sample size of 200 (for comparison with African-American group) and sample size of 500 (for comparison with Hispanic-American group).

For the case of using SAT score as the sole predictor for GPA, empirical sampling distributions for sample sizes of 200 and 500 were constructed respectively for the intercept and the slope by bootstrapping 1,000 resamples from the white group data. The

sample intercepts and slopes of African-American and Hispanic-American groups were compared with the bootstrapped sampling distributions. If, compared with the bootstrapped sampling distribution, the sample estimates for African-American and Hispanic-American groups' regression lines exceeded the random sampling variation, then the null hypothesis would be considered untenable. Consequently, it would be concluded that there was a real difference in slopes or intercepts between the majority and the minority groups, and the use of a common prediction line for all the groups would result in predictive bias for certain group(s).

When both SAT and relative high school rank were used to predict the cumulative GPA, a common regression plane was fitted to all the groups, and the residuals for each group were obtained. Estimated empirical sampling distributions (for sample sizes of 200 and 500) of the mean residuals were constructed by bootstrapping 1,000 resamples from the white group residuals. The sample mean residuals for the African-American and Hispanic-American groups were then examined relative to the bootstrapped sampling distributions. If the mean residuals of the minority groups exceeded the random sampling variation, it would be considered evidence of predictive bias.

Implementation of Conventional and Bootstrap Approaches

The conventional approach to the case of one predictor (SAT score) was implemented through the regression procedure (PROC

REG) under the Statistical Analysis System (SAS, Version 6.08 for Microsoft Window). Group membership was represented by two effect coding variables (1, 0 and -1) in order to test for intercept and slope differences among the groups. The testing procedures as suggested by Kerlinger and Pedhazur (1973) and Pedhazur (1982) were used to test for potential slope and intercept differences.

For bootstrap approach, all bootstrap resampling (sampling with replacement) and calculations were accomplished by using the Interactive Matrix Language (PROC IML) under the Statistical Analysis System (SAS). Random sampling with replacement for bootstrap procedure was accomplished by utilizing the random number generator for uniform distribution (RANUNI under SAS) to generate a vector of random numbers ($m \times 1$ dimension, with m being the desired bootstrap sample size). Each element of the vector was randomly and independently generated (to accomplish the feature of "with replacement" required by bootstrap), and was constrained to be integers between 1 and n inclusive, with n being the number of observations of the white group. This vector of integers was then used as the index numbers to draw row vectors (samples) from the matrix of original white group data.

Results and Discussions

Results of Conventional Statistical Testing

The results from the conventional statistical significance testing approach and the bootstrap approach were both presented,

with those from the conventional approach presented first. When only SAT score used as the sole predictor for GPA, three regression models, with each nested under the previous one, were fitted to the data, and the hypothesis for common slope and that for common intercept were statistically tested through nested regression models. The three regression models were as follows:

1) Full Model: the three groups differ in both the slope and the intercept;
2) Reduced Model 1: the three groups have the same slope, but differ in intercept;
3) Reduced Model 2: the three groups have the same slope and the same intercept.

The approach for testing for extra sum-of-squares was used:

$$F_{(df_r - df_f,\ df_f)} = \frac{SSE_{reduced} - SSE_{full}}{df_{reduced} - df_{full}} \div \frac{SSE_{full}}{df_{full}}$$

and the following two hypotheses were tested sequentially: a) H_0: the three groups share the same slope but have different intercepts; and b) H_0: the three groups share both the same slope and the same intercept.

Table 1 presents the statistical significance testing results for the nested regression models. It is seen that the null hypothesis of common slope can be retained, since the use of a common slope did not increase the Sum-of-Squares for error significantly. The null hypothesis for a common intercept (conducted when the common slope hypothesis was retained), on the other hand, could be rejected due to the statistically

significant increase in the Sum-of-Squares for error. Further statistical testing (not presented in Table 1) indicated that the intercept for White group was higher statistically than those of Hispanic-American and African-American groups, while no statistical difference existed between African-American and Hispanic-American groups.

According to the results of this approach, since the groups shared the same slope but differed in intercept, the use of a common regression line (Reduced Model 2: same slope and same intercept) would result in predictive bias. A close examination for the direction of bias revealed that the bias would underpredict for White group GPA, and over-predict for the two minority groups' GPA.

When both SAT score and the relative high school rank were used for GPA prediction, a common regression plane was fitted to the data regardless of ethnic group membership. The residuals of the three groups were compared through ANOVA technique to determine if there was any statistically significant difference in the mean residuals of the three groups which could indicate predictive bias for certain groups. Table 2 presents the descriptive statistics of the residuals for the three ethnic groups, and Table 3 presents both the ANOVA analysis results for testing the residual differences among the groups, and the results of post hoc multiple comparisons (Tukey's Studentized Range Test).

Insert Table 2 About Here

Insert Table 3 About Here

Table 2 and Table 3 show that the mean residuals among the ethnic groups were statistically different. Post hoc comparisons (Tukey Studentized Range Test with Type I experimentwise error rate controlled) revealed that the difference occurred between White group on one hand, and African-American and Hispanic-American groups on the other. No statistical difference was detected between African-American and Hispanic-American groups. Since the residual is obtained by subtracting the predicted GPA from the observed GPA ($e = Y - \hat{Y}$), negative residual indicates over-prediction for GPA, while positive residual indicates under-prediction for GPA. Again, this residual analysis for the case of two predictors indicates under-prediction for the white group and over-prediction for both the African-American and Hispanic-American groups.

Results from the Bootstrap Approach

Instead of relying on statistical significance testing, bootstrap approach empirically estimated the sampling distributions based on data of White group, and the sample estimates for African-American and Hispanic-American groups were examined relative to this sampling distributions. Sample

estimates exceeding sampling variation would be considered as evidence to reject the null hypothesis for a common slope or a common intercept for different groups.

Figure 2 presents the bootstrapped sampling distributions (for sample size of 200 and 500 respectively) for the intercept (SAT used as the sole predictor for GPA), and the locations of the sample estimates for African-American and Hispanic-American groups. It is obvious that the sample estimates of the two minority groups are located well within the empirically bootstrapped sampling distributions from the white group data. Put in other words, if we were to draw one random sample of 200 from the white group, the probability would be about 0.20 that we would observe an intercept equal to or smaller than 0.252, the one obtained by the African-American sample. Similar conclusion could be drawn for the Hispanic-American sample. Since the sample estimates of the minority groups did not exceed what would be expected from sampling variation, we could only conclude that the three groups had the same regression intercept. This conclusion is contrary to that based on the statistical significance testing approach (Table 1) which concluded that the three groups differed in their intercepts, and the white group had statistically higher intercept than those of the African-American and Hispanic-American groups.

Figure 3 presents the bootstrapped sampling distributions (for sample size of 200 and 500 respectively) for the regression slope (SAT used as the sole predictor for GPA), and the locations

of the sample estimates from the African-American and Hispanic-American groups. Similar to the previous intercept situation, Figure 3 shows that the sample estimates for the regression slope of the minority groups are well within the sampling distributions bootstrapped from the white group data, thus the null hypothesis for a common slope was retained. This was in agreement with the conclusion from the previous statistical significance testing approach.

Figure 4 presents the case when two predictors (SAT score and Relative High School Rank) were used to predict GPA. For this analysis, a common regression plane was fitted to all the data regardless of group membership, and the residuals for the three groups were obtained. The sampling distributions for mean residuals were constructed by bootstrapping the white group data, and the sample estimates from the minority groups were examined relative the bootstrapped sampling distributions. It is clear that the sample estimates of mean residuals for the two minority groups are well beyond the empirically bootstrapped sampling distributions. In other words, for GPA prediction, after fitting a common regression plane for all the three groups, if we were to draw a random sample of 200 out of the white group, the probability would be almost zero that we would observe a mean residual as small as that obtained by the African-American sample. The probability to observe the value of mean residual obtained by the Hispanic-American sample would be even smaller. Since the sample estimates for the two minority groups well

exceeded what could be expected from random sampling variation, it was concluded that there were real differences among the residuals of the three groups if a common regression plane was used for all the three groups.

Residual difference among the groups indicated the existence of predictive bias for certain groups. A closer examination revealed that the predictive bias was against the white group (positive mean residual for underprediction) but in favor of both African-American and Hispanic-American groups (negative mean residuals for overprediction), a conclusion consistent with that based on the statistical significance testing approach.

Conclusions

The two different approaches for investigating potential predictive bias produced somewhat different results. When only SAT was used for first year cumulative GPA prediction, the statistical significance approach concluded that the groups shared a common regression slope, but their regression intercepts differed. Because of the intercept differences for the groups, the use of a common regression model would result in predictive bias which was against the white group but in favor of the African-American and Hispanic-American groups.

The bootstrap approach, on the other hand, concluded that the three groups shared the same slope (consistent with the statistical significance testing approach), and that the intercept differences of the groups did not exceed sampling

variation, either (contrary to the statistical significance testing results). Based on the locations of the minority groups' intercept estimates relative to the bootstrapped sampling distributions, the conclusion from the bootstrap approach seems to be more plausible for the data. It is very possible that the large sample size utilized in significance testing approach (see Table 1) may have contributed to the statistical significance.

When two predictors were used (SAT and relative high school rank), the two approaches came to the same conclusion: the mean residuals of the three groups are different, indicating predictive bias for certain groups. The predictive bias was revealed to be in favor of the two minority groups but against the white group.

The paper demonstrated the feasibility of applying bootstrap technique to the investigation of potential predictive bias in college GPA prediction. Since bootstrap procedure is more empirically grounded, some problems associated with the conventional statistical significance testing approach, such as sample size (too large or too small?) and the validity of theoretical assumptions for particular data sets, may be avoided.

References

Bentler, P. M. (1992). EQS Structural Equations Program Manual. Los Angeles, CA: BMDP Statistical Software.

Bickel, P. & Freedman, D. (1981). Some asymptotic theory for the bootstrap. Annals of Statistics, 9, 1196-1217.

Carver, R.P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Cleary, T.A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 5, 115-124.

Daniel, L. G. (1992). Bootstrap methods in the principal component case. Paper presented at the Annual Meeting of the American Educational REsearch Association (San Francisco, CA, April 20-24, 1992). ERIC #: ED 346135.

Diaconis, P. & Efron, B. (1983). Computer-intensive methods in statistics. Scientific American, May, 116-130.

Efron, B. (1979). Bootstrap methods: Another look at the jacknife. The Annals of Statistics, 7, 1-26.

Freedman, D. A. (1981). Bootstrapping regression models. Annals of Statistics, 9, 1218-1228.

Goldman, R. D. & Richards, R. (1972). The SAT prediction of grades for Mexican-American versus Anglo-American students at the University of California, Riverside. ERIC#: ED 088904.

Hand, C. A. & Prather, J. E. (1985). The predictive validity of Scholastic Aptitude Test scores for minority college students. Paper presented at the Annual Meeting of teh American Educational REsearch Association. March 31-April 4, 1985. ERIC #: ED261093.

Harris, D. J. & Kolen, M. J. (1988). Bootstrap and traditional standard errors of the point-biserial. Educational and Psychological Measurement, 48, 43-51.

Harris, D. J. & Kolen, M. J. (1989). Examining the stability of Angoff's delta item bias statistic using bootstrap. Educational and Psychological Measurement, 49, 81-87.

Hunter, J. E., Schmidt, F. L. & Rauschenberger, J. (1984). Methodological, statistical, and ethical issues in the study of bias in psychological tests. In C. R. Reynolds and R. T. Brown (Eds.), Perspectives on bias in mental testing. New

York: Plenum Press.

Jensen, A. R. (1980). Bias in mental testing. New York: Free Press.

Kerlinger, F.N., & Pedhazur, E. J. (1973). Multiple regression in behavioral research. New York: Holt, Rinehart & Winston.

Lambert, Z. V., Wildt, A. R., & Durand, R. M. (1990). Assessing sampling variation relative to number-of-factor criteria. Educational and Psychological Measurement, 50, 33-48.

Lambert, Z. V., Wildt, A. R., & Durand, R. M. (1991). Approximating confidence interval for factor loadings. Multivariate Behaviroal Research, 26, 421-434.

Lunneborg, C.E. (1987). Bootstrap applications for the behavioral sciences. Seattle: University of Washington.

Lunneborg, C.E. (1990). [Review of computer intensive methods for testing hypotheses]. Educational and Psychological Measurement, 50, 441-445.

Mendoza, J. L., Hart, D. E., & Powell, A. (1991). A bootstrap confidence interval based on a correlation corrected for range restriction. Multivariate Behavioral Research, 26, 255-269.

Pedhazur, E. J. (1982). Multiple regression in behavioral research: Explanation and prediction. New York: Holt, Rinehart and Winston.

Potthoff, R. F. (1966). Statistical aspects of the problem of biases in psychological tests (Institute of Statistics Mimeo Series No. 479). Chapel Hill: Department of Statistics, University of North Carolina.

Reynolds, C.R. (1980). An examination of bias in a preschool battery across race and sex. Journal of Educational Measurement, 17, 137-146.

Reynolds, C.R. (1982). Methods for detecting construct and predictive bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD: The Johns Hopkins University Press.

Rogosa, D. (1980). Comparing nonparallel regression lines. Psychological Bulletin, 88, 307-321.

Sue, S. & Abe, J. (1988). Predictors of academic achievement among Asian American and white students. New York, N.Y.:

College Entrance Examination Board.

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-6.

Thompson, B. (1992a). Exploring the replicability of a study's results: Bootstrap statistics for the multivariate case. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC No. ED 344895)

Thompson, B. (1992b). DISCTRA: A computer program that computes bootstrap resampling estimates of descriptive discriminant analysis function and structure coefficients and group centroids. Educational and Psychological Measurement, 52, 905-911.

Thompson, B. (1993). The use statistical significance tests in research: Bootstrap and other alternatives. The Journal of Experimental Education, ? ?

Willson, V. L. (1985). Analysis of interactions in research. In C. R. Reynolds and V. L. Willson (Eds.), Methodological and statistical advances in the study of individual differences. New York: Plenum Press.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. The Annals of Statistics, 14, 1261-1294.

Table 1: Testing for Nested Regression Models for Common Slope and Common Intercept

Model	SSE	DF	Results	Regression Models
Full Model	2547.02	4946		B[a]: GPA = 0.251 + 0.203(SAT)[b] H: GPA = 0.422 + 0.185(SAT) W: GPA = 0.612 + 0.183(SAT)
Reduced Model 1	2547.16	4948	H_0: same slope, different intercept $F_{(2,4946)}=0.135$ $p > 0.10$	B: GPA = 0.427 + 0.184(SAT) H: GPA = 0.429 + 0.184(SAT) W: GPA = 0.604 + 0.184(SAT)
Reduced Model 2	2564.62	4950	H_0: Same slope, same intercept $F_{(2,4948)}=16.96$ $p < 0.01$	(Common Regression Line:) GPA = 0.485 + 0.193(SAT)

a: B: African-American; H: Hispanic-American; W: White
b: The unit for SAT score in the models is 100.

Table 2: Descriptive Statistics for Residuals of the Three Groups (SAT and High School Rank to Predict GPA)

Group	n	Median	Mean	STD
African-American	195	-0.1608	-0.1736	0.7362
Hispanic-American	509	-0.1637	-0.1675	0.7289
White	4248	0.0897	0.0280	0.6583

Table 3: ANOVA Test for Equal Mean Residuals for the Three Groups (SAT and High School Rank to Predict GPA)

Source	DF	SS	MS	F	P
Ethnic Group	2	23.49	11.74	26.23	0.001
Error	4949	2215.70	0.45		
Total	4951	2239.19			

Tukey's Grouping (Type I Experimentwise Error Rate Controlled)

	Mean	Groupings
White	0.02803	Different
Hispanic-American	-0.16747	Same
African-American	-0.17356	Same

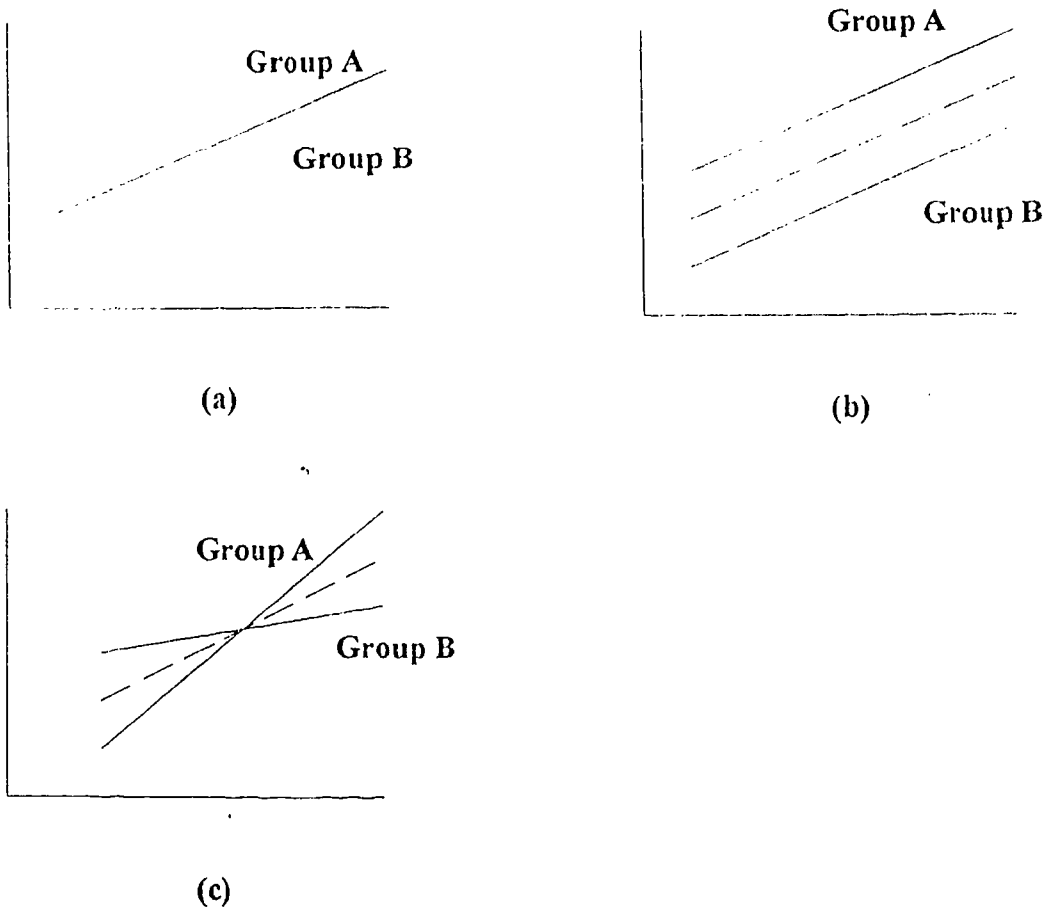
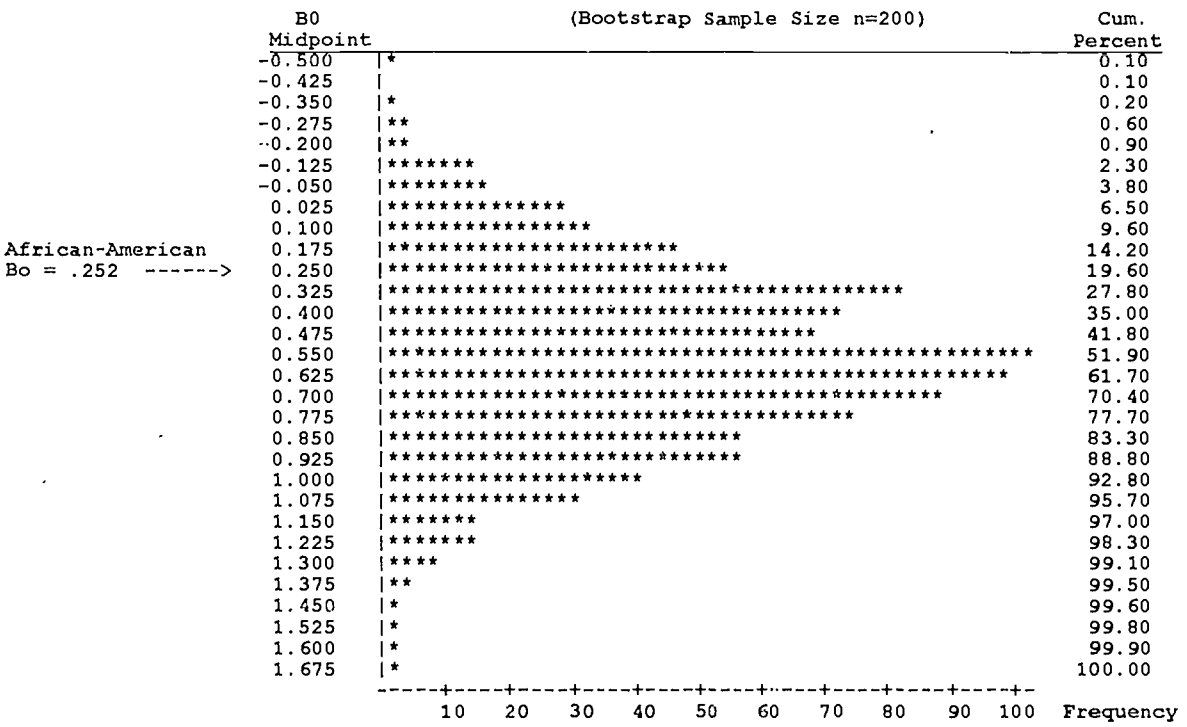


Figure 1: Three Situations Related to Criterion Prediction



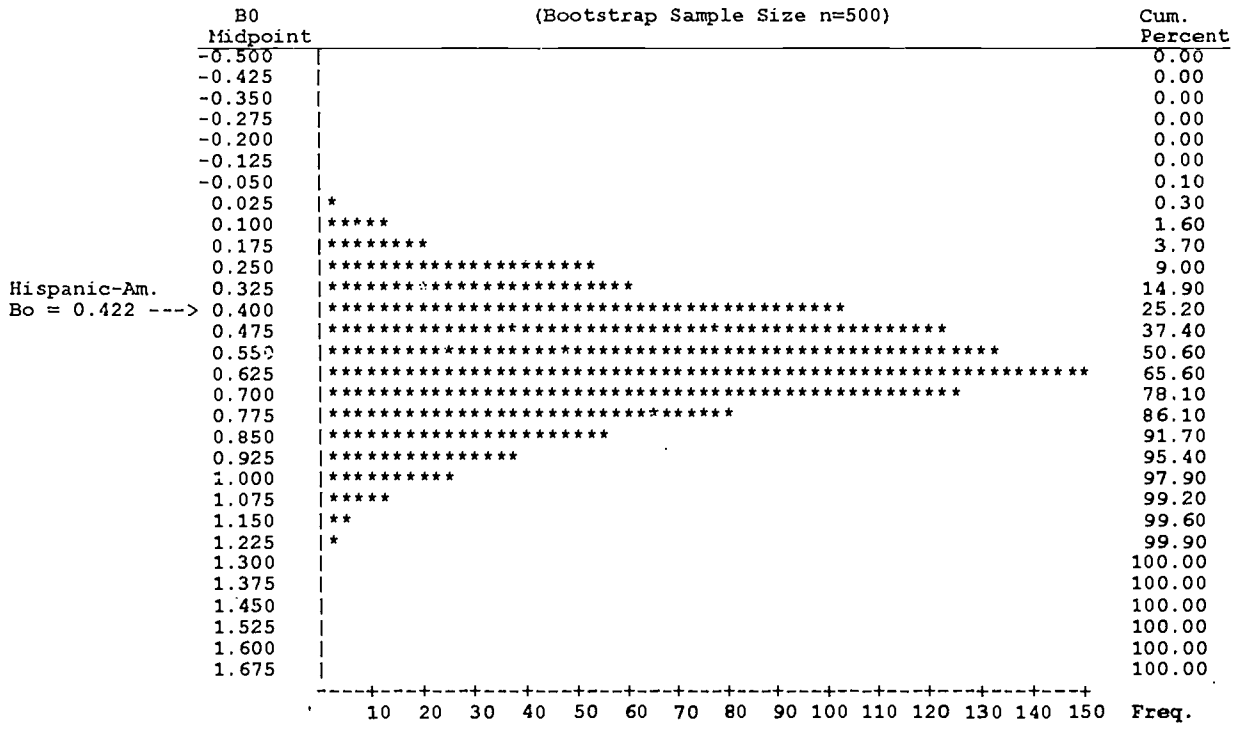
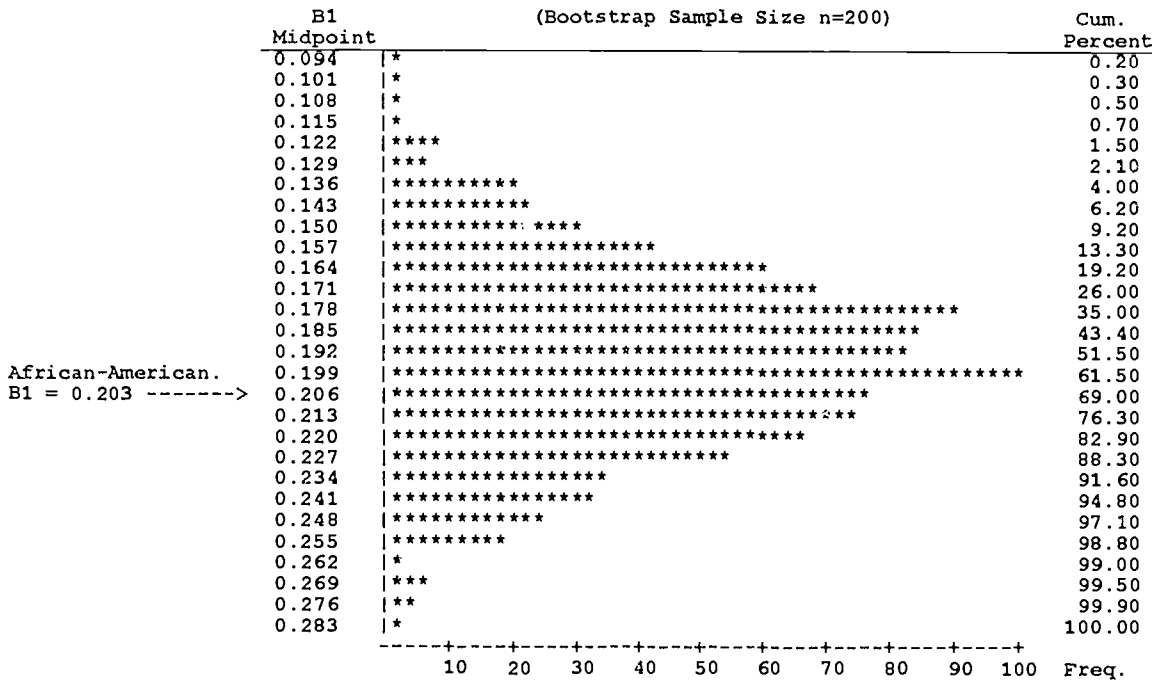


Figure 2: Minority Groups' Sample estimates for Intercept versus Sampling Distributions Bootstrapped from the White Group (SAT as Sole Predictor)



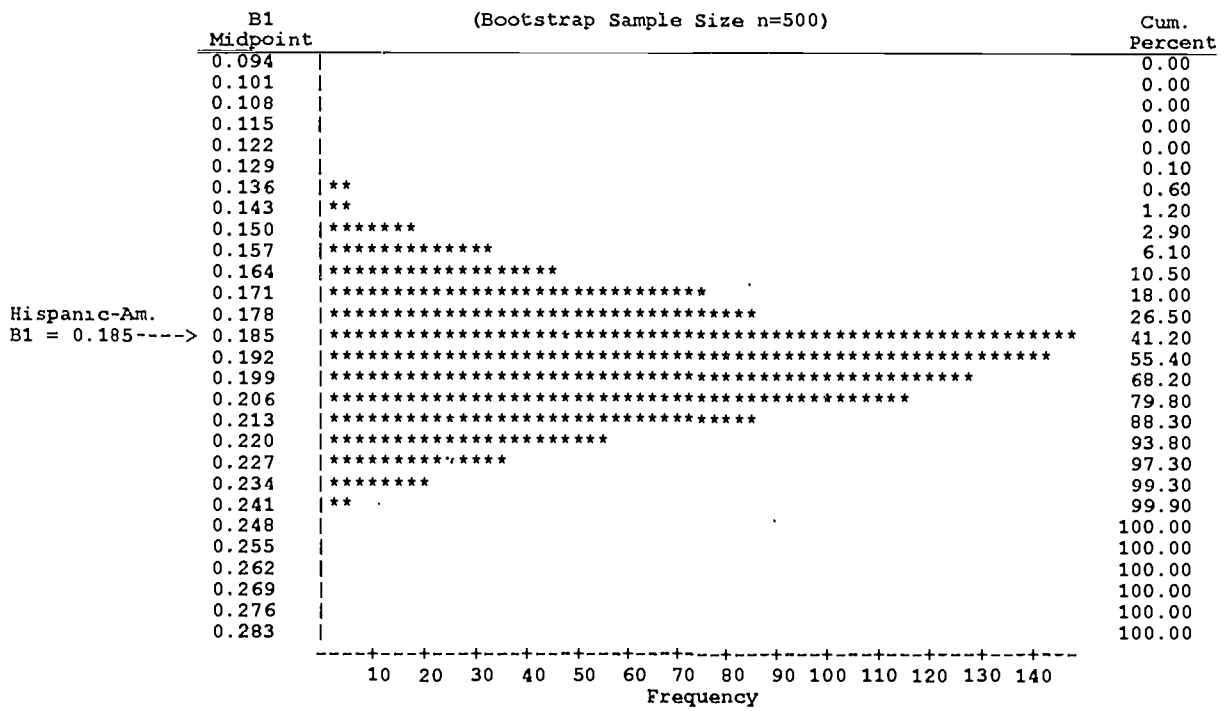
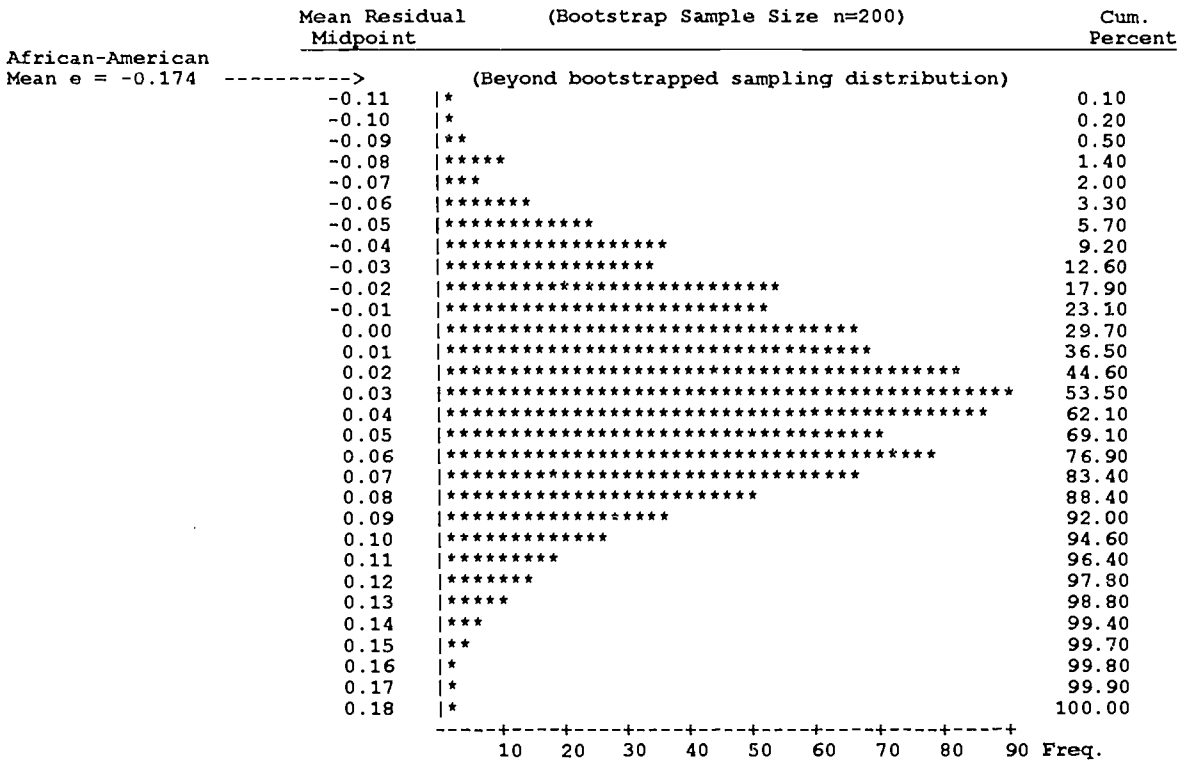


Figure 3: Minority Groups' Sample Estimates for Slope versus Sampling Distributions from the White Group (SAT as the Sole Predictor)



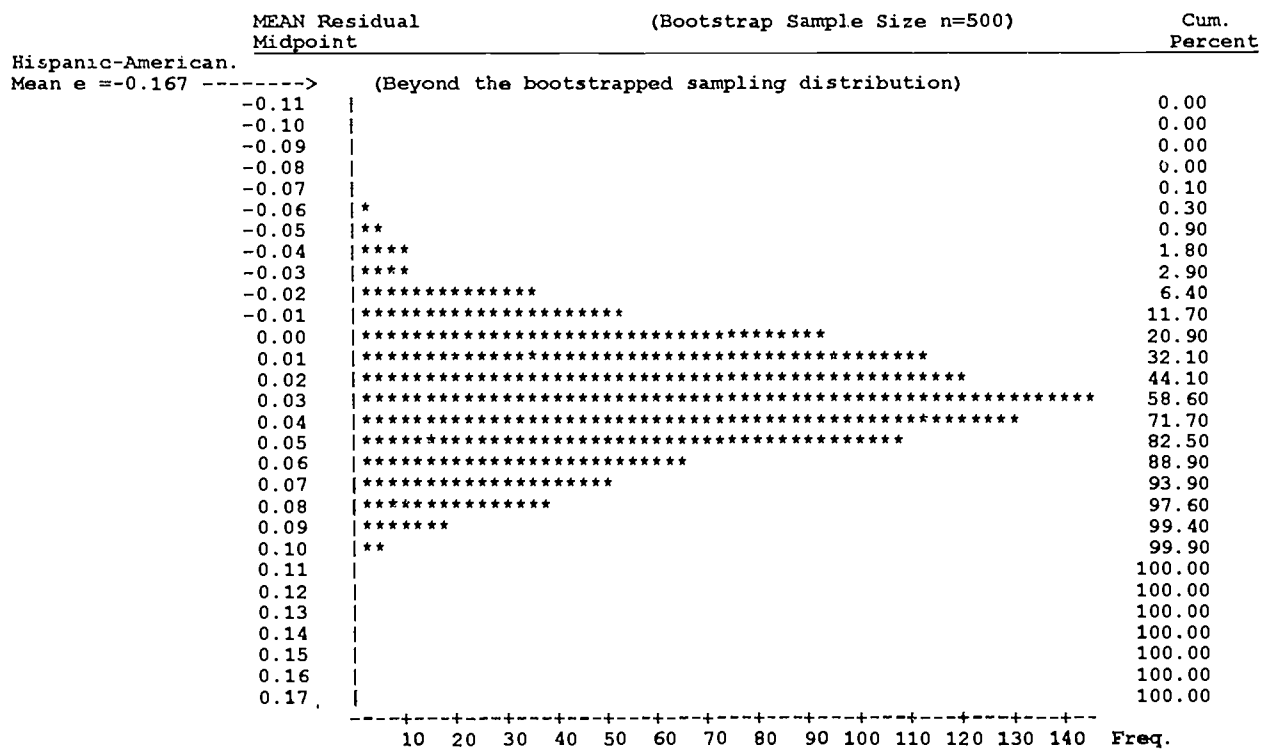


Figure 4: Minority Groups' Sample Estimates for the Mean GPA Residuals versus Sampling Distributions from White Group (Both SAT and High School Rank Used as Predictors)